A bayesian network approach to model local dependencies among SNPs

Raphaël Mourad[1], Christine Sinoquet[2], Philippe Leray[1]

[1] Laboratoire d'Informatique Nantes Atlantique, UMR CNRS 6241
Ecole Polytechnique de l'Universitï¿½ de Nantes, la Chantrerie - rue Christian Pauc - BP 50609, 44306 Nantes Cedex 3 France
raphael.mourad,philippe.leray@univ-nantes.fr
[2] Laboratoire d'Informatique Nantes Atlantique, UMR CNRS 6241
Universitï¿½ de Nantes, 2 rue de la Houssiniï¿½re, BP 92208, 44322 Nantes Cedex 03, France
christine.sinoquet@univ-nantes.fr

Abstract: *In this preliminary work, we investigate a method to model linkage disequilibrium among SNPs (Single Nucleotide Polymorphisms) in the genome. The genetic data such as SNPs is characterized by a typical block-like structure along the genome. Graphical models such as bayesian networks can provide a fine and biologically relevant modeling of dependencies for both haplotypical and genotypical SNP data. We applied a MWST-based algorithm (Maximum Weighted Spanning Tree) to construct a bayesian network, relying on the underlying local dependencies.*

Keywords: Linkage disequilibrium, bayesian networks, haplotype blocks, genetic association studies.

1 Foreword

This research work was presented at MODGRAPH 2009, Satellite Meeting of JOBIM 2009 (Probabilistic graphical models for integration of complex data and discovery of causal models in biology), Nantes, 8 june [http://www.lina.univ-nantes.fr/conf/modgraph/]

2 Introduction

Genome wide association studies (GWAS) address the localisation and identification of causal mutations responsible for common genetic diseases. At the present time, such studies fail to identify the combination of loci involved in multifactorial diseases. Those studies exploit the linkage disequilibrium existing between SNPs. Linkage disequilibrium (LD) is the non-random association among alleles, for two or more loci, on the genome [5]. LD is mesured on haplotypical data (phased genotypes) and not on genotypical data. LD exists within chromosomes, among SNPs which lay in close vicinity. Beside such local dependencies, LD can exist for distant SNPs, on the same chromosome or for SNPs located on two different chromosomes, thus defining global dependencies. The HapMap project revealed the block-like structure of linkage disequilibrium on human genetic data [4].

In the context of a GWAS dedicated to a group of multifactorial heart diseases, the mitral valvular dystrophies, our purpose consists in designing a new approach based on bayesian networks, to model

local dependencies among SNPs. Bayesian networks (BN) show several advantages for this purpose: learning models from data, integration of heterogeneous data, possibility to analyse a vast amount of variables and ability to take into account expert knowledge [6]. Several authors worked on this subject: Verzilli *et al.* used Markov networks to model LD and perform a genetic association study [8]. Nefian proposed hierarchical bayesian networks with hidden variables to learn SNP dependencies [7]. However, we think those works do not take into account the complex structure of SNP dependencies: blocks of dependencies of various lengths and dependencies between neighbouring blocks. This complex structure of SNP dependencies justifies the modeling through bayesian networks.

3 Methods

First, we want to explore how local SNP dependencies are structured. For this purpose, we studied a genotypical sequence corresponding to a 81kb-long region located on chromosome 1 of the human genome (see HapMap project [http://www.hapmap.org/]). We used Gevalt software [2], based on the expectation-maximization algorithm, to compute LD plots and to infer haplotypes and haplotype blocks (haploblocks).

Second, we used the GeNIe SMILE© software to learn the optimal network structure with a greedy search-based algorithm. The greedy search consists in optimizing a score when exploring the graph space [6]. The analysis of the optimal forest seems to be an appropriate approach to simply investigate the sequence of dependencies among SNPs. We compelled the BN to be a forest by allowing just a single parent per node.

Third, learning BN structures using MWST-based algorithms has been shown efficient and scalable [3]. Usually, the MWST algorithm is used to build the optimal tree [6]. Thus, we modified this algorithm to find an optimal forest. The MWST algorithm successively adds to the tree under construction the edges with the best weight, *i.e.* : $W(X_1, X_2) = score(X_1, \Pi_{X_1} = X_2) - score(X_1, \Pi_{X_1} = \emptyset)$, where $score(X, \Pi_X)$ is the score of the node X knowing it has a parent Π_X. To identify an optimal forest, we stop adding edges when $W(X_1, X_2)<0$, (*i.e.* there is no score increase). We analyzed a sequence of 1735 kb at the beginning of the chromosome 1, which provides 489 SNPs.

4 Results

4.1 Block-like structure of genetic data

Results are shown in figures 1a and 1b. Figure 1a shows 5 haploblocks on the sequence. When comparing to LD plot on figure 1b, we observe that haploblocks capture a good amount of dependencies between SNPs. But we also notice some dependencies between haploblocks (surrounded by a dotted line).

4.2 Optimal forest through a greedy search-based algorithm

We have modeled the dependencies among SNPs constructing a bayesian network from haplotypical data (figure 1c). The same process has been applied with genotypical data (figure 1d). Examining the haplotypical data, we see that some haploblocks or parts of haploblocks are present in the graph (highlighted with a dotted line). For example, SNPs from haploblocks 2 are found directly dependent. We also see that the SNPs from haploblock 5 are quasi directly dependent through SNP 59. Regarding the bayesian network relative to genotypical data, we find similar dependencies (surrounded by a dotted line). Thus, the LD on haplotypes leads to block-wise dependencies at the genotypical level.

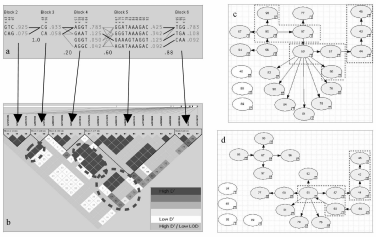

Figure 1. a) 5 haploblocks infered from a 81kb sequence on 90 individuals. b) LD plot. The darker a square, the higher the LD is between the two SNPs concerned. Haploblocks are highlighted in black. c) Bayesian network on haplotypical data. d) Bayesian network on genotypical data.

4.3 Optimal forest through a MWST-based algorithm

We have implemented our algorithm in ProBT®[1], a C++ library specialized in bayesian networks. We successively ran the algorithm with different scores : BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion). We found similar results with BIC and AIC. The running time averages 10 minutes. The obtained forest shows 12 connected components with an average of 40.75 SNPs per connected component.

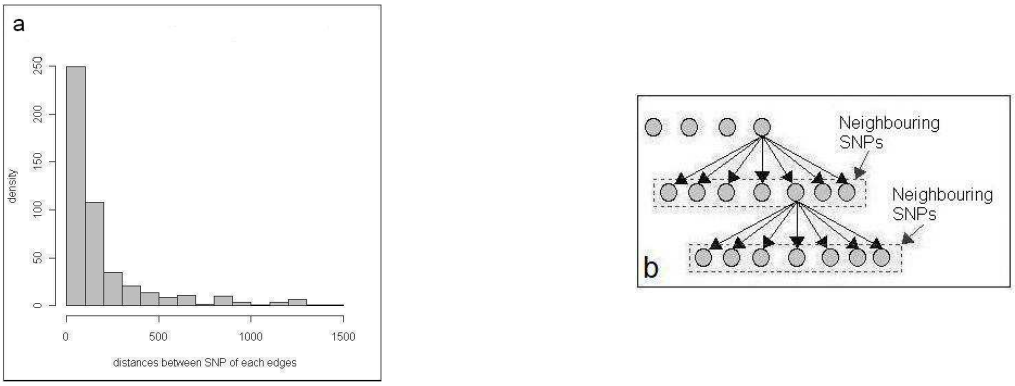

Figure 2. a) Number of edges in the network as a function of the distance between SNPs. b) Diagram of the obtained forest.

We calculated the distribution of the distances between the SNPs of each edge (figure 2a). We observe that the decay of strength of the dependencies between SNPs with distance is well modeled by the forest. Unfortunately, this method suffers from the fact that most of the SNPs (98%) are present on the same connected component. Indeed, each of the 11 first SNPs of the sequence are pairwise independent and the 478 other SNPs form only one connected component. The results are presented through a simplistic diagram in figure 2b. Unexpectedly, we do not obtain a block-wise structure such as previously revealed by greedy search, on data of smaller size. Using MWST algorithm, we report a big connected component where most of neighbouring SNPs are connected to the same parent nodes (surrounded by a dotted line). Moreover, we increased the threshold to stop adding edges in order to get a larger number of connected components: we obtained similar results. At the present time, this method does not allow the modeling of blocks of LD.

5 Conclusion and outlooks

First, we began to investigate the applicability of the MWST algorithm to model local SNP dependencies. For this purpose, this algorithm has been adapted to the identification of a forest. A straightforward application of this algorithm fails to find connected components which mimic the LD blocks in DNA. Future work will focus on fixing or penalizing the number of children of a node during the forest construction. Second, we think bayesian networks can represent an useful tool for geneticists. The first aim is to allow to explore the linkage disequilibrium between SNPs in a more readable way than does a LD plot. The final aim is using the BN as the structure to be investigated for the identification of association between combinations of markers and a multifactorial disease.

Acknowledgements

The first author is thankful to the BIL project that provides support for this current research work. The BIL project is dedicated to enhance research in the field of Bioinformatics, in the "Pays de la Loire" Region.

References

[1] J. M. Ahuactzin, K. Mekhnacha, E. Mazer, P. Bessière, A brief introduction to Bayesian modeling with ProBT, user's manual, 2005.

[2] O. Davidovich, G. Kimmel and R. Shamir, GEVALT: An integrated software tool for genotype analysis, *BMC Bioinformatics*, 8:1-8, 2007.

[3] O. François and P. Leray, ï¿½Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens, *RJCIA 2003*, Laval,167-180, 2003.

[4] The International HapMap Consortium, A haplotype map of the human genome, *Nature*, 437:1299-1320, 2005.

[5] D. L. Hartl and A. G. Clark, Principles of population genetics, Sinauer associates, Inc., edition 3, 1997.

[6] P. Naïm, P. H. Wuillemin, P. Leray, O. Pourret and A. Becker, Réseaux bayésiens, Eyrolles, edition 3, pp 131-153, 2007.

[7] A. V. Nefian, Learning SNP dependencies using embedded Bayesian networks, *IEEE Computational Systems, Bioinformatics Conference 2006*, 1-6, 2006.

[8] C. J. Verzilli, N. Stallard and J. C. Whittaker, Bayesian graphical models for genomewide association studies, *The american journal of human genetics*, 79:100-112, 2006.